Development of Text-To-Speech System for Latvian

Kārlis Goba
Tilde
Vienības gatve 75a, Riga, LV1004
Latvia
karlis.goba@tilde.lv

Andrejs Vasiļjevs
Tilde
Vienības gatve 75a, Riga, LV1004
Latvia
andrejs.vasiljevs@tilde.lv

Abstract

This paper describes the development of the first text-to-speech (TTS) synthesizer for Latvian language. It provides an overview of the project background and describes the general approach, the choices and particular implementation aspects of the principal TTS components: NLP, prosody and waveform generation. A novelty for waveform synthesis is the combination of corpus-based unit selection methods with traditional diphone synthesis. We conclude that the proposed combination of rather simple language models and synthesis methods yields a cost effective TTS synthesizer of adequate quality.

1 Introduction

This paper describes the development of the first text-to-speech (TTS) synthesis system for Latvian.

The Latvian language spoken by 1.6 million people is the only official language in Latvia and one of the working languages of European Union. Despite the important role of Latvian, until now there was no text-to-speech synthesizer for this language. As a result, there are no applications in use providing Latvian speech capabilities.

The population group with the most acute need for speech enabled technologies are visually impaired people. TTS is an essential technology enabling them to use computer applications, browse the internet and communicate via e-mail. Some of them are advanced computer users using English TTS, but majority do not have sufficient English skills. Attempts to use English TTS for reading and preparing Latvian texts have failed due to principal differences in Latvian and English pronunciation. Latvian text pronounced by English TTS is practically incomprehensible even by the most tolerant and striving users.

In late 1990s and the beginning of this century, a few experiments on Latvian TTS were carried out by the Institute of Informatics and Mathematics of the University of Latvia. These were experiments of speech generation by concatenation of individually recorded phonemes. It became clear quite soon that this approach cannot lead to human-like speech and the experimental system was never completed.

Ilze Auzina (2005) has summarized many of the aspects that need to be accounted for in speech synthesis of Latvian. Auzina proposes models for syllable boundary detection and a framework for grapheme-to-phoneme conversion.

Juris Grigorjevs has carried out research on Latvian speech analysis at Department of Philology of University of Latvia. In this research Grigorjevs (2005) carried out experiments on Latvian vowel generation using formant synthesis.

There has been an attempt to create a Latvian adaptation of the *WinTalker* system developed by a Czech company called *RosaSoft*. The pronunciation generated by pilot model was better than Latvian texts pronounced by non-Latvian TTS but still of a very low quality and barely recognizable. For this reason it was not accepted by users and was not further developed.

The current development project of Latvian TTS synthesizer was started in 2005. The project is carried out as part of a European Commission funded programme to facilitate accessibility for impaired people.

The following sections describe the general approach and the motivation behind the decisions made while creating the text-to-speech synthesizer.

Joakim Nivre, Heiki-Jaan Kaalep, Kadri Muischnek and Mare Koit (Eds.)
NODALIDA 2007 Conference Proceedings, pp. 67–72

2 TTS overview

The primary purpose of Latvian TTS is to address the needs of visually impaired people using computers in the Latvian language environment – browsing Latvian internet, reading and creating Latvian documents, enabling e-mail and chat communication in Latvian.

The requirements of this project include natural sounding text-to-speech synthesis in Latvian that can be integrated with screen-reading accessibility software for visually impaired people. The basic requirements of TTS engines, like the possibility to change voice pitch and speech rate, as well as support for user pronunciation dictionaries are included. Language processing within the TTS engine has to be robust and functional in order to provide stable and consistent output in different usage environments.

The system architecture covers the traditional text-to-speech transformation, performing text normalization, grapheme-to-phoneme conversion, prosody generation, and waveform synthesis.

Latvian in general can be considered a *phonetic language* – a language with relatively simple relationship between orthography and phonology as defined by Huang et al. (2001).

From the TTS synthesis perspective, Latvian has several specific properties:

- Short and long vowels and consonants
- Largely phonetic orthography
- Highly inflected language
- Uniform stress pattern
- Lexical syllable tones

These properties have to be taken into account in text normalization, grapheme-to-phoneme conversion and prosody generation. The following subsections describe their impact.

2.1 Language processing

Language processing consists of text normalization and the subsequent conversion to narrow phonetic transcription.

As a first step, text containing words, abbreviations, numbers, punctuation and other symbols is transformed to normalized orthography.

Latvian has a rich inflectional system. Nouns, adverbs, verbs and participles take different forms depending on gender, number, case, degree, definiteness, mood, tense and person. The constituents of sentence are required to be in agreement between each other. This influences the disambiguation of abbreviated text elements during text normalization.

Some Latvian abbreviations have fixed representations while others should be represented in the inflected form depending on the context:

u.tml. *un tamlīdzīgi* (‘and so on’)
u.c. *un citi / citiem / …* (‘and others’)

The fixed representations are included in text processing rules. However, currently no processing is done to determine the inflection of inflected abbreviations.

Appropriate inflectional form should be also determined while transcribing measurement units and numbers:

5 g *pieci grami / piecu gramu / …*

Here simple contextual rules are used. The ending of the next word is used to determine the inflectional form of the numeral.

Pronunciation of acronyms depends on linguistic traditions. Acronyms traditionally are read letter by letter, though in some cases they are pronounced phonetically, in Latvian or in their original language:

ASV */ā es vē/*
PVN */pē vē en/*
LTV7 */el tē vē septiņi/*
NATO */nato/*
SIA */siā/*
KNAB */knab/*
UNESCO */junesko/*
Reuters */roiters/*

The pronunciation of acronyms has to be included in the user pronunciation dictionary.

In the Latvian orthography, the letters *e*, *ē* and *o* are homographs and can denote different phonetic values depending on the lexical properties of the word. In some cases the lexical information is not sufficient to distinguish:

vēlu */ve:lu/* (verb ‘to roll’)
vēlu */væ:lu/* (verb ‘to wish’ or adverb ‘late’)
robots */ruobuots/* (adjective ‘serrate’)
robots */robots/* (noun ‘a robot’)
aerobs */aero:bs/* (adjective ‘aerobic’)

In these cases, part-of-speech tagging or morphologic analysis may be used for phonetic disambiguation. For efficiency reasons and since such homographs are not very frequent in Latvian, we include only the statistically most frequent forms in disambiguation rules.

Latvian orthography tries to retain the morphologic structure of words as much as possible, while observing a number of pronunciation rules. These rules have to be accounted for also in the grapheme-to-phoneme conversion, e.g. regres-

sive assimilation between subsequent voiced and unvoiced consonants, which also occurs across word borders.

Language processing is performed by over 1,000 regular expression search-and-replace rules.

2.2 Prosody modelling

Latvian has several properties of tonal languages. Each word has an associated stress placement and tone pattern. Both the stress and the tone pattern are distinctive lexical features, e.g. the minimal pairs:

nékur ((he)does not make fire)
nekúr (nowhere)
mãja (house)
mājа ((he) waved)

Syllables with a long nucleus (long vowel, diphthong or vowel and syllabic sonorant) have one of the three distinct tones present in Latvian. The tones are lexical features of morphologic constituents (the root and optional prefixes, suffixes and endings).

However, quantitative research on the usage of tones in Latvian is lacking. Laua (1997) describes the three tones in a qualitative way:

Tone	Description
Stretched	Pitch is rising steadily to high
Falling	Pitch is slowly falling from high
Broken	Pitch and intensity are rising until the break, when the intensity suddenly drops and resumes after the break

In modern spoken Latvian, two of the three long syllable tones merge together depending on regional dialect of the speaker. The three-tone system, being the richest and the oldest, is currently preserved only in certain regions (Laua 1997). Experiments suggest that it is acceptable to follow this tendency and to model only two distinct syllable tone patterns: stretched and non-stretched (combined rising and broken tone).

The tone pattern is most distinctive in stressed syllables. Experiments suggest that it is acceptable to model the tone of unstressed syllables with a neutral pitch contour. Stressed syllable tone has a lexical function in modern Latvian, while the tone distinction in unstressed syllables has a minor influence on understanding and perceiving of speech.

The syllable stress in Latvian is expressed by emphasizing the tonal contour and lengthening of the stressed syllable. In general, the syllable stress falls on the first syllable. Exceptions include a fixed list of words that have historically merged together (e.g. *labvákar* < *lábu vákaru*), as well as the superlative degree of adjectives and adverbs (e.g. *vislábākajam*).

The Fujisaki pitch model has been successfully adapted for many languages, including such tonal languages as Swedish and Chinese (Fujisaki 2004). Experiments with Fujisaki model showed that the syllable tone accents in Latvian can be sufficiently modelled with one or two accent commands near the nuclei of stressed syllables.

To model syllable and phrase level stress, discrete prosodic events are inserted in the narrow phonetic transcription. This processing is rule-based. To obtain F0 contour, the prosodic events are converted to accent and pitch commands.

The prosodic events are located at stressed syllable nuclei and the boundaries of prosodic phrase. Prosodic phrases are determined in a simplified way as indicated by text punctuation.

The vowel length is lexical and is marked in orthography with macron diacritics. The consonant length is denoted as double consonants. The length of plosives and fricatives is also influenced by the phonetic context.

Duration is also modelled in discrete steps. Special symbols denoting relative increase or decrease in duration are inserted in the narrow phonetic transcription according to manually written rules. The rules include only the well-known regular phenomena of the Latvian language: the lengthening of unvoiced plosives between two short vowels, the lengthening of stressed syllables and the shortening of short final syllables. Other prosodic factors (phonetic context, structure and position of words, metric feet and phrases) are not taken into account.

2.3 Waveform synthesis

According to Morais and Violaro (2005), corpus-based synthesis approach is the dominant trend in this decade in speech synthesis, which provides high naturalness, accuracy and intelligibility.

In corpus-based synthesis, pre-recorded speech units are concatenated and transformed to produce speech. At runtime, appropriate acoustic patterns and the prosody of a sentence are superimposed during concatenation by means of digital signal processing techniques.

Drawbacks of the corpus-based synthesis are the high development costs and the relatively high memory and processing requirements for

running the system. Corpus-based synthesis using a large number of larger recorded units like sentences, words, phrases and morphemes, may produce higher quality speech, but requires a lot of processing power and memory for unit storage.

According to the traditional approach, only one recorded speech unit for each diphone is stored in the diphone synthesis system, and it is presumed that diphones are context-independent. This implies that each diphone has to be carefully selected and evaluated against other diphones that might precede or follow it in order to minimize the possible discontinuities.However, this is a very time-consuming task.Moreover, diphones are influenced by the context and the presumption does not quite hold true.

To improve the quality of Latvian TTS, it was decided to store multiple variations of most diphones and to select the appropriate variant during speech synthesis.

Initially, several variations of each diphone in different contexts were recorded to increase the represented variation of vowels and consonants.

Then, several subsequent diphones were marked in frequently occurring words, including weekdays, months, frequent country and city names etc. During synthesis, these subsequent diphones have minimal join costs during diphone selection, thus allowing use of effectively larger speech units.

The size of the phone set in Latvian is quite disputable and there is no agreement on this in literature. 30 consonant allophones, 2 glides and 6 short and 6 long vowels can be identified, as well as numerous diphthongs. Including all these phones in a phone set significantly increases the size of diphone set above 2000 diphones.

For diphone synthesis, the phone set has to be acoustically representative, i.e. covering the various spectrally steady regions (phone centres) for speech production. Thus experiments were done to decide on the possible size reduction of the acoustic inventory.

Grigorjevs (2005) has shown that the spectral properties of phonemically short and long vowels in Latvian are not distinctive, thus long vowels can be perceived as only differing in duration.

During the development of TTS, experiments of time-scale modification of recorded speech were carried out. Results suggested neither vowels nor unvoiced plosives and consonant sonorants of different length and duration show considerable spectral differences. This allows simplifying the diphone set by treating long and short phonemes uniformly.

The diphone set was further simplified and reduced in size by treating affricates as consecutive plosives and fricatives (/ts/ /tš/ /dz/ /dž/) and treating diphthongs as two consecutive vowels. To cover the possible allophonic and spectral variation of these component phones, multiple occurrences of diphones containing the subject phones in different contexts were recorded.

After the reduction, the acoustic phone set consists of 29 phones including 6 vowel phones, 22 consonant phones and a silence phone. That gives 841 phone pairs, of which about 750 diphones are possible in Latvian.

The speech material for diphones includes one or several words for each possible phone combination. These words are wrapped in carrier sentences to provide natural speech flow.

Several professional speakers were tested to select the most appropriate voice. Recorded sentences were phonetically segmented and manual selection of diphones was made from the recorded material.

The total size of the diphone database is ~2,200 diphones, on average containing 3 variations of each phone pair. Each diphone is divided into overlapping pitch-synchronous windows and the respective LPC coefficients and the residual signal are stored in the diphone database.

The LPC analysis is used for two purposes. First, it allows separating the source (excitation) from the filter (formants) within the limits of linear model. Applying pitch and duration modification to the LPC residual allows introducing less distortion in modified speech. Second, the LPC coefficients are used for evaluating the spectral distance between two diphones during diphone candidate selection at synthesis.

During synthesis, the diphone candidates are selected by dynamic programming algorithm that minimizes the cumulative sum of unit join costs. Currently only spectral distance estimate is used for join costs. The selected diphones are then pitch-synchronously concatenated and the LPC residual signal is modified to the target pitch and duration by overlap-add method. The waveform is generated by LPC synthesis.

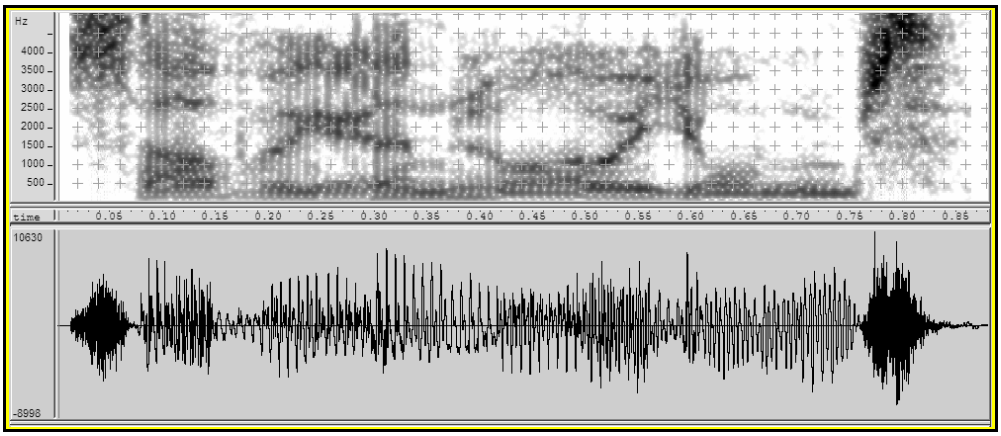

Figure 1. The spectrogram and waveform of a synthesized word /*savienuojums*/ (connection). Random diphone variations are chosen. Note the discontinuities.

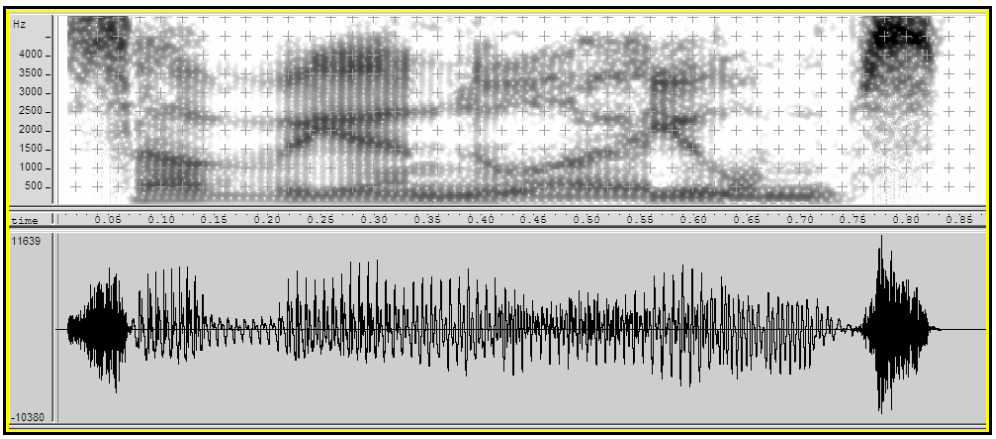

Figure 2. The spectrogram and waveform of a synthesized word /*savienuojums*/ (connection). LPC coefficient-based diphone selection is used. Note the smooth concatenation at diphthong */ie/* and the discontinuities at diphthong /*uo*/.

3 Conclusions and further work

The Latvian TTS system described in this paper currently is being beta tested in real usage scenarios. Feedback from the first users is very positive. They characterize generated speech as natural-sounding, with correct Latvian pronunciation of majority of phrases and efficient work even on relatively old systems (200 MHz Pentium processor). The project demonstrates applicability of diphone synthesis in combination with such speech quality improvements as usage of multiple diphone variations and LPC residual modification.

The combined approach of diphone synthesis and unit selection provides a good compromise between the speed and the effectiveness of speech synthesis, and the quality of the produced speech:

- Increasing the variety of vowel and consonant pronunciation,
- Decreasing the spectral discontinuity between adjacent diphones (for comparison, see Figure 1 and Figure 2),
- Enabling "reconstruction" of longer speech units what were adjacent in speech material used for diphone extraction.

However, Latvian prosody is not yet fully developed. The simplified phrase prosody model shows good results with relatively short sentences. Such sentences are frequent when screen reading software is used by visually impaired people for interface with computer. However, it is more difficult to follow such synthetic voice for longer and more complex sentences.

Further work will concentrate on development of the speech rhythm and intonation model, evaluation of the resulting system and the implemented improvements in comparison to the classical diphone synthesis, and integration of the speech synthesizer with different application scenarios.

References

Ilze Auzina. 2005. *Computer Modelling of Latvian Pronunciation: synopsis of doctoral thesis*. Riga, Latvia.

Hiroya Fujisaki, 2004. *Information, Prosody, and Modelling with Emphasis on Tonal Features of Speech*. Speech Prosody 2004, Nara, Japan.

Juris Grigorjevs. 2005. *Acoustic and Auditory Characteristics of Latvian Vowel System: synopsis of doctoral thesis.* Riga, Latvia.

Xuedong Huang, Alex Acero, Hsiao-Wuen Hon. 2001. *Spoken Language Processing. A Guide to Theory, Algorithm, and System Development.* Prentice Hall.

Alise Laua. 1997. *Latviešu literārās valodas fonētika (Phonetics of literary Latvian), 4th edition*. Riga.

Edmilson Morais and Fįbio Violaro. 2005. *Data-Driven Text-to-Speech Synthesis.* XXII Simpósio Brasileiro de Telecomunicações, Campinas, Brazil.